DOCUMENT RESUME

ED 144 941 TM 005 519

AUTHOR Williams, John D.
TITLE Testing and the Testing Industry: A Third View.
INSTITUTION North Dakota Univ., Grand Forks. Center for Teaching and Learning.
SPONS AGENCY Rockefeller Bros. Fund, New York, N.Y.
PUB DATE Dec 75
NOTE 49p.
AVAILABLE FROM Vito Perrone, Center for Teaching and Learning, University of North Dakota, Grand Forks, North Dakota 58202 ($2.00)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; Intelligence; Intelligence Tests; Nature Nurture Controversy; *Standardized Tests; *Test Bias; Test Construction; *Testing Problems; Test Validity; Test Wiseness
IDENTIFIERS Alternatives to Standardized Testing; Jensen (Arthur R); Testing Industry

ABSTRACT

Different viewpoints regarding educational testing are described. While some people advocate continuing reliance upon standardized tests, others favor the discontinuation of such achievement and intelligence tests. The author recommends a moderate view somewhere between these two extremes. Problems associated with standardized testing in the schools include the tendency to evaluate educational programs according to their effect upon test scores, the discontinuation of worthwhile programs that do not contribute to improved test scores, and the amount of time required to administer test batteries. Problems associated with incorrect interpretation of intelligence tests, and comments on Jensen's article regarding the relative impact of genetic and environmental factors on intelligence are also discussed. The growth of the testing industry is described, and alternatives to present methods of testing are suggested. The author recommends that those groups most involved in the testing process should begin, in a meaningful way, to consider the uses and abuses of testing. (GDC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

North Dakota Study Group on Evaluation

ED144941

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

John D. Williams

TESTING AND THE TESTING INDUSTRY: A THIRD VIEW

TM005 519

Copyright © 1975 by John Williams

First published in 1976

North Dakota Study Group
on Evaluation, c/o Vito Perrone,
Center for Teaching & Learning
University of North Dakota
Grand Forks, N.D. 58201

Library of Congress Catalogue
Card number: 75-46138

Printed by University of
North Dakota Press

A grant from the Rockefeller Brothers Fund
makes possible publication of this series

Editor: Arthur Tobier

In November 1972, educators from several parts of the United States met at the University of North Dakota to discuss some common concerns about the narrow accountability ethos that had begun to dominate schools and to share what many believed to be more sensible means of both documenting and assessing children's learning. Subsequent meetings, much sharing of evaluation information, and financial and moral support from the Rockefeller Brothers Fund have all contributed to keeping together what is now called the North Dakota Study Group on Evaluation. A major goal of the Study Group, beyond support for individual participants and programs, is to provide materials for teachers, parents, school administrators and governmental decision-makers (within State Education Agencies and the U.S. Office of Education) that might encourage re-examination of a range of evaluation issues and perspectives about schools and schooling.

Towards this end, the Study Group has initiated a continuing series of monographs, of which this paper is one. Over time, the series will include material on, among other things, children's thinking, children's language, teacher support systems, inservice training, the school's relationship to the larger community. The intent is that these papers be taken not as final statements--a new ideology, but as working papers, written by people who are acting on, not just thinking about, these problems, whose implications need an active and considered response.

Vito Perrone, Dean
Center for Teaching & Learning,
University of North Dakota

Contents

	Introduction	1
1	Jensenism and Anti-Jensenism	5
2	A Genetic Approach to Intellectual Inheritance	14
3	Looking at Tests and Items on Tests: Hoffman Revisited	17
4	Use of Intelligence Testing in the Schools	20
5	Use of Standardized Achievement Tests in the Schools	23
6	Why Do Schools Administer Standardized Tests?	25
7	Do We Need This Big an Educational Testing Industry?	27
8	An Interim Solution: Teaching Students to Take Tests	30
9	A More Lasting Solution: Constructing Locally-Useful Criterion-Referenced Tests	32
10	An Alternative Solution: Constructing Criterion-Referenced Work Samples	34
11	A Final Statement	37
	References	39

Introduction

The debate regarding standardized testing in general, and intelligence testing in particular, has quietly risen to a commanding position on the agenda of all those parties trying to influence the policy of American education. In the process, the debate has trimmed out and clearly marked its boundaries.

One edge of the debate is centered on the age-old heredity-environment controversy; this aspect of the debate was renewed by Jensen (1969), with Shockley (1971) and Herrnstein (1971) in a supporting cast role. Their collective view on the issue of heredity and environment as it affects intelligence might be characterized by saying heredity predominates and accounts for perhaps 80 percent of human variability in intelligence test scores. Their protagonists (Sitgreaves, 1961, Kagan, 1969, Lewontin, 1973, Fehr, 1969 and Morris, 1972, to name but a few) are not so easily classified as to the causative agents of human intelligence, but they do agree that it is not due to heredity alone.

Another edge of the debate centers upon test validity, either as it applies to particular items or as it applies to testing subgroups not originally measured in the norming of the test. A most provocative explication of the inappropriateness and inaccuracy of some test items that appear on widely available tests was made by Hoffman (1962). More recently, the March/April 1975 issue of *Principal* was devoted to "The Myth of Measurability" and included articles criticizing the items that make up several tests, including the *Lorge-Thorndike Intelligence Tests*, the *Wechsler Intelligence Scale for Children*, the *Otis-Lenon Mental Ability Test*, and *The Iowa Tests of Basic Skills*, to name but a few. A subsequent issue (July/August) was devoted to "The Scoring of Children: Standardized Testing in America." This second issue was even more critical of the testing industry, calling into question almost every one of its practices. The opposite view is expressed by the test-constructing industry and its defenders.

Another edge of the debate centers upon test usage: How are the schools using (or going to use) the data available on each child? If the school is to develop learning settings for "non-mainstream" children (classes for the learning disabled or classes for the gifted), how important should be the role played by standardized test-

John D. Williams, Professor, Center for Teaching and Learning, The University of North Dakota. His teaching area is applied statistics. His articles have appeared in *Multivariate Behavioral Research*, *The Journal of Experimental Education*, *The Journal of Educational Research*, *Educational and Psychological Measurement*, *The Journal of Genetic Psychology*, *The Journal of Psychology*, *Psychological Reports*, *Behavioral Science*, *Perceptual and Motor Skills*, *Research in Higher Education*, *Multiple Linear Regression Viewpoints*, and several other professional journals.

ing? One point of view is that if the labelling process overrepresents or underrepresents an identifiable subgroup (females, blacks, Spanish-surnamed, lower socioeconomic status, left-handed persons or Catholics), then the test is discriminatory and should be abandoned. Padilla and Garza (1975) point out that Spanish-surnamed children have a 2-to-2 1/2 times greater chance of being involved in classes for slow learners than others in Texas and California. They contend that "IQ" tests are an inadequate measure for the Spanish-surnamed because they do not sample the cultural and linguistic experiences likely to be a part of the child's background. They also are among several professional educators calling for a moratorium on intelligence testing of minority children. Far from trying to defend itself, or even turning a deaf ear to the criticism, the test industry has seen this as an opportunity to sell more and newer tests. In fact, a plethora of tests have been recently rewritten for specific minorities; Samuda (1975) has an extensive annotated bibliograpny of tests for minority students.

UNDERLYING PREMISES

Perhaps much of the debate alluded to in the preceding section is due to very different underlying premises of the various participants. To take two "straw men" at each node of the continuum, a staunch defender of standardized testing might reason that the tests, for the most part, have been rigorously standardized in field testing and that a viable product has been achieved; the corresponding "anti-tester," at the other end of the continuum, might reason that tests that are made up of such fallible items as those that occur on typical tests can mean no more in their sum than the contribution of each item; if the items are as questionable as they seem to be, then the scores are most likely meaningless. Further, if the tests empirically show themselves to be used to place people into slower learning situations and an identifiable subgroup has more than its "share" of people so identified, this is *prima facie* evidence that the test is discriminatory toward the affected group. It is clear, then, that people with such discrepant belief systems will likely look at the same set of data and draw entirely different conclusions.

The premises of the present paper are few. First, in regard to the teaching and learning of students, it is felt that students should be in learning situations that allow them to maximize their learning potential while minimizing the interference of non-learning situations. No specific learning environment is prescribed; indeed, if Hunt (1971) is right, students should be matched with learning environments that best fit their conceptual functioning; what might be a maximally functioning environment at one stage of the student's learn-

ing may be inappropriate at a later stage.

A second premise is that the conventional wisdom of any group is always open to question. As the conventional wisdom of any group has a way of being modified over time (perhaps because not all members of a group accept the conventional wisdom), typifying the "tester" or the "anti-tester" is always time-dimensioned and hence inaccurate shortly after attempting to typify, or "stereotype," that conventional wisdom. For example, those involved in intelligence testing surely no longer include in their thinking such bromides as (a) intelligence is fixed and unchanging; (b) without considerable intelligence (say, intelligence scores of 130 and above as measured by *Wechsler's Adult Intelligence Scale*), such attainments as graduate degrees are impossible (or at least highly improbable); or (c) crime and lower intelligence are inseparable.

Interestingly, the critics of the intelligence testing movement are guilty of at least one major *faux-pas* when they continually refer to "IQ" testing. The IQ, or "Intelligence Quotient," has passed from the lexicon of most psychometricians who engage in intelligence testing. There were too many inconsistencies in attempting to arrive at a measure of intelligence by the fabled formula, $IQ = \frac{MA}{CA}(100)$ where:

IQ = the intelligence quotient,
MA = the mental age, and
CA = chronological age.*

One problem is that the gradients of learning are not smooth enough to allow the IQ scores to have a sufficient degree of predictability in longitudinal studies. The constructors of the *Stanford-Binet* test coined the term "Deviation Quotient" and developed norms at each age as an alternative to the difficulties of the traditional IQ. Other intelligence test constructors (and users) have opted for the term "intelligence score" or even "academic aptitude test."

The point is that test constructors and allied psychometric personnel are not particularly impressed by condemnations of a term (IQ) that has been out of general usage for more than a decade. To be fair, the term IQ continues in the vocabulary of Jensen, Shockley and Herrnstein. Perhaps the psychometric community is being inadequately represented in this debate?

The conventional wisdom of the "anti-testers" is not nearly so closely delineated as it is for the "testers"; those opposed to standardized testing have almost as many positions on testing as there are people in opposition to the testing movement. For example, one might say that the present conventional wisdom asks that a moratorium be made on intelligence testing in public schools. Others might hold that this moratorium be made on utilizing data from some very specific subpopulation. In between these two points occur many clearly staked-out

*One recent author who continues to use this concept is McCall (1975). In an attempt to explain the intelligence and heredity concept at a fairly elementary level, McCall opted for a simplistic definition of an intelligence test score. Perhaps part of the rationale for this is that the term "IQ" is ubiquitous in the jargon of non-psychometrists. By capitalizing on the familiarity of this term, McCall may have felt that the reader's interest might more likely be captivated, and, in the process, make him more fully informed on the intelligence-heredity issue. Unfortunately, the perceptive reader may see the "flaw" in the definition of intelligence, and thus reject the point of McCall's writing.

positions.

A third premise holds that alternative explanations and/or research methodologies may show a very different light on a given topic. For example, in regard to the heredity-environment issue, it seems logical to look outside the usual educationist attempts at research and at least become aware of any research relating to intellectual functioning from a geneticist's point of view.

It should be clear from these premises, then, that the point of view being developed is a dynamic viewpoint; as more evidence accumulates, from whatever its source, some rethinking is necessary. So is it necessary that no evidence be discounted because its source fails to produce the proper credentials, either educationally or philosophically. Whether it is presented by a black woman with only an eighth grade education or by a person such as Shockley, who is neither a psychologist nor an educationist, the evidence is not to be rejected with *ad hominem* arguments.

Such an *ad hominem* argument appears to have been made in an otherwise reasoned defense of the testing movement by Ebel (1975, p. 83):

> Education is blessed with a great many capable and dedicated teachers and administrators. But the profession also has its share of mediocrity and of false messiahs. It is from the latter group that the loudest protests are heard against tests and testing.

To be placed in the camp of "mediocrity and of false messiahs" simply because one holds a view different from the "expert" is not very comforting. Green (1975), I think, did well to point out that the majority of those who oppose testing do so from the point of view that minority group interests are often violated by abuses of the tests.

1

Jensenism and Anti-Jensenism

Perhaps no scholarly publication has attracted more attention than Jensen's lengthy article in the *Harvard Educational Review*. The reaction against him has been so strong in some quarters that it seems compelling to find out its cause. At the American Educational Research Association Convention in Chicago, in April 1972, where Jensen was scheduled to address a research audience regarding his findings, Chicago city teachers not only picketed the convention, urging researchers not to attend his presentation, but several stormed the hall, and among other indignities shown to Dr. Jensen, his notes were stolen. Moreover, demonstrators caused so much disruption that his presentation was cancelled. For such an unusual amount of attention to be foisted upon a researcher (albeit negative attention), it behooves other researchers to at least find out about the squabble and the research that generated it. What, in fact, has been the case?

In the past decade, perhaps a handful of educational research studies have captured the public's, or rather the media's, attention. Besides Jensen's work, Coleman *et al.'s* (1966) monumental analysis of data conducted at the request of the then President Johnson is notable; Rosenthal and Jacobson's (1968) study focused attention on teachers' attitudes toward students and helped to popularize the term "self-fulfilling prophecy"; Jencks's (1972) reanalysis of Coleman's data was an attempt to assess, to some degree, Jensen's findings on the importance of heredity. In each case, very few people, including researchers, appear to have reached back to the original source materials for their information, but instead have relied on the media or condensations and/or rebuttals by others. This is understandable to an extent, as each of these research efforts is lengthy and would require detailed examination. This is particularly true of Jencks's study; if one were not a statistician, understanding Jencks's results would almost seem to be out of the question. But of those professionals who have responded in print regarding Jensen's work, the criticisms are not nearly as strong as the media might have led us to believe. For example, take the reaction of a noted "anti-tester," Banesh Hoffman, in a recent interview:

> When Jensen wrote his paper, I read about it in the New York Times and I was highly incensed. I was ready to go out on the hustings and write a letter to the Times, and all that. Then I thought, before I do that, I better read his paper. I did so and was surprised to find that it was really a serious, honest paper. I think, however, that Jensen has not realized one important thing that is very hard to measure, and that is the effect environment has on a person. For instance, if you are a black child, you can sense the hatred that is focused on you; you realize that if you do anything good, no one is going to like it coming from a black kid. So you have children growing up in an atmosphere of oppression and hatred. I don't think that Jensen realized how terrible that is, how it can stunt the intellectual and emotional growth of a person. (Banesh Hoffman, interviewed by Houts, 1975, p. 36.)

A review of Jensen's article is in order.

A REVIEW OF JENSEN'S ARTICLE

The article begins, ironically, with the point on which his evidence is weakest: compensatory education has failed. He does point out that some of the early efforts at compensatory education (Project Headstart and related activities) have often had less than glowing successes. But he does not consider the perspective of those in the community served by the various programs. To a black, Chicano, or Indian, the compensatory education programs of the 1960s must have seemed something less than a panacea for all their needs. (In fact, that black or Chicano or Indian might have remarked, "Isn't it funny that all the massive amounts of money available always go into white hands? The administrators of the program are white; the teachers are white; and all the materials bought are bought from white businessmen. Don't they trust us?")

Next, Jensen ventures into a standard textbook explanation of the nature of intelligence (24 pages). In it, he considers several conceptions of intelligence, but opts for a pragmatic solution: measure certain kinds of behavior, look at their relationship to other phenomena, and see if the relationships make any sense. While some might have grave misgivings about not pinning down the entity (intelligence), Jensen's approach is fairly standard in psychological research. Jensen quotes several studies regarding the correlation of intelligence measures and indices of socio-economic status (the r's range from .42 to .71).

Jensen also enters a discussion of genotype (the genetic make-up) and phenotype (an observable or measurable characteristic of an organism). The square of the

correlation between the genotype and the phenotype is called the hereditability of that trait. He discusses the distribution of intelligence scores. He fails, however, to go into the construction of intelligence tests, a shortcoming I find singularly perplexing.*

Then Jensen considers the inheritance of intelligence from a genetic viewpoint. First, he reviews the work of Burt (1958, 1966), which was an attempt to sort out the proportion of the variance in intelligence tests relating to genetic and environmental factors.** Second,

---

*While the construction of intelligence tests is necessarily heavily statistical by nature, much of the construction process uses methods ordinarily taught in a first course. Wechsler's is quite illuminating in this regard both because the test is highly respected, and because the manual does an excellent job of describing the norming process.

Wechsler has two scales for children, *Wechsler's Intelligence Scale for Children* (WISC), ages 5-15, and the *Wechsler Preschool and Primary Scale of Intelligence* (WPPSI), ages 4-6 1/2, and a scale for adults, *Wechsler's Adult Intelligence Scale* (WAIS). The WISC includes six performance tasks and six verbal tasks. The WAIS omits one performance task and is, of course, at a more difficult level. The one omitted task is the mazes task. The reason for its omission says quite a bit about intelligence test construction; among children, there are no significant sex differences on any of the 12 tasks. For adults, men score significantly higher on the mazes task; because an assumption is made that there is no sex difference on any of the 12 tasks, those tasks that do show a difference are discarded. (One might ask, as has been done by Bane (1974), why don't we get rid of items that show *race* discrimination, thus rendering as academic the whole issue brought up by Jensen?)

Each task that is completed on the Wechsler tests yields a scaled score (not necessarily a point for each correct response); the scaled scores are then added, and their sum is transformed to a new scaled score, the intelligence score. Originally, Wechsler standardized his data so that the mean would be 100, the standard deviation would be 15, and there would be no sex differences. Further, the data were scaled so that a normal distribution would result. Nevertheless, some interesting anomalies occur in parts of the test. For example, no more credit is given for getting all the items correct than if the longest digit span is missed.

** Burt's data appear to be the seeds of Jensen's concept that intelligence is 80 percent determined by genetic factors. While it is difficult to criticize an author on a second-hand basis, it appears Burt used an hierarchical model (see Cohen, 1968) that insured the pre-

he reviews the various kinship studies, which, in turn, were earlier reviewed by Erlenmeyer-Kimling and Jarvik (1963). He also reviews the twins studies, from which he draws several points: first, that correlations in intelligence scores of identical twins are systematically higher than for fraternal twins; second, that identical twins who have been separated tend to have more similar test scores than fraternal twins raised together.*

* Kamin (1974) notes the lack of scientific controls in most of these studies.

Finally, Jensen reviews studies regarding environmental correlates of intelligence, getting to the core issue only on page 78 of a 123-page article. In fact, had the first two pages of the article been omitted, and had the article ended on page 77, Jensen would probably still be an obscure educational psychologist. But after page 77, Jensen gingerly rationalizes why it is at least interesting to investigate racial differences in intelligence, as racial differences in many other aspects of human life have been investigated. As Jensen passes the point of no return (discussing racial differences), a grossly oversimplified synopsis of his reasoning might be useful. First, he attempts to show that intelligence is to some degree (he says 80 percent) hereditary; second, he attempts to show that blacks score an average 15 points below whites on intelligence tests; third, he attempts to show that compensatory education has failed. His evidence for the first point is rather substantial. His evidence on the second point is somewhat weaker, but given the misgivings, at least the relation of the evidence is usable. The third point has little documentation in the article; it seems to be the crux of his presentation, however.

A little discussed point from this part of Jensen's article relates to the American Indian. The American Indian is said to be ". . . by far the most environmentally disadvantaged group . . .," yet their scores on intelligence tests put them only seven to eight points below the Caucasian mean. Indeed, on at least one test (Dennis, 1942), American Indians have been shown to score *higher* than Caucasians. In fact, a group of Hopi had a mean intelligence score of 124, clearly superior to most Caucasian samples. Jensen argues that if environment is causal to the black-Caucasian differences in intelligence, then how can environment logically be related to

---

dominance of genetic factors. Had environment been used as the first-ordered variable, a very different estimate of the importance of heredity would probably have been found. Kamin (1974) thoroughly attacks the Burt studies on the grounds that their lack of adequate reporting and apparent assessment process of adult intelligence on a non-testing base reduces them to an interesting commentary on psychological research in the middle third of the twentieth century, clearly lacking in terms of today's standards.

the Indian differences?

Jensen then considers a eugenics argument that does have a data base. Apparently, upper and middle class blacks have a substantially lower birth rate than their Caucasian counterparts. The reverse is true of lower class blacks and Caucasians. To whatever degree social class is correlated to intelligence, and that relationship, in turn, is correlated to heredity, Jensen conjectures that future studies would show larger discrepancies between the two races on measures of intelligence.

The remainder of Jensen's work is concerned with the various studies of compensatory education. The important point in the later portion of his article is that, in separating intelligence into two levels (Level One, Associative Learning, and Level Two, Conceptual Learning), schools tend to emphasize conceptual learning, while those at the lower ability levels achieve a greater degree of success with associative learning. Jensen would encourage a much greater emphasis on associative learning for lower ability students.

REVIEWS OF JENSEN'S ARTICLE IN SUBSEQUENT ISSUES OF THE *HARVARD EDUCATIONAL REVIEW*

In the following two issues, the *Harvard Educational Review* published several critiques of Jensen's paper, written before and after the paper had become a rallying flag. Kagan (1969) suggested that Jensen's arguments were not compelling, but only suggestive. Clearly his conclusions regarding compensatory education, Kagan wrote, were inappropriate on several grounds; it is illogical to use current evaluations to dismiss all possible compensatory programs. J. McV. Hunt's (1969) criticisms were astonishingly mild; his greatest criticism focused on Jensen's having stated "compensatory education has been tried and it appears to have failed." Hunt saw that as a half-truth placed at the beginning of the paper for its dramatic effect. Crow (1969), a geneticist, agreed for the most part with Jensen's analysis, but pointed out the limitations of the mathematical assumptions, the sample size, and lack of evidence regarding changes in the environment that had not yet been tried.

Bereiter (1969, p. 310) stated, "My own view of the future of individual differences and their social consequences is even less optimistic than Dr. Jensen's. The hereditability of intelligence is unquestionably high . . . with further social progress, its reliability can only increase . . ." Conceptually complex machines such as computers that are understandable to only a small percent of the populace will inevitably magnify individual differences in ability, Bereiter wrote. Elkind (1969) described a Piagetian conception of intelligence; he also related that his evidence indicates that ". . . *the longer we delay formal instruction, up to certain limits, the greater the period of plasticity and the higher the ulti-*

mate level of achievement." Thus, nursery schools (and the compensatory programs for pre-school children) might better try to create experiences that have immediate value for the child.

Cronbach (1969) agreed to some extent with Jensen's findings, but found the Levels I and II distinction in intellectual functioning to be an oversimplification. Brazziel (1969) took umbrage with Jensen regarding the application of similar research to the blacks in the South. His concern was that racists will use the article in an attempt to continue or reestablish as much segregation as they can legally achieve.

Clearly, the earlier criticisms, written prior to the publication of Jensen's article, were not nearly so marked as they were reported in the public press. The reviews that appeared in the second post-Jensen paper issue of the *Review*, when the media had taken it up, were somewhat stronger.

What might be characterized as one of the better critiques of Jensen's statistical methodologies was made by Light and Smith (1969). They showed that almost 9 points of the 15-point difference in mean intelligence scores between Caucasians and Negroes can be attributed to the disproportionate distribution of socio-economic status as it relates to "race." They further argued that, if an unusual interaction pattern is present, the entire 15-point difference can be taken into account. Unfortunately, they presented no empirical evidence for the unusual interaction pattern. If they had, its social consequences would seem to be unacceptable; in some intelligence levels it would be more beneficial to have a lower standard of living, rather than a higher level. Few families would seem to want to move into abject poverty for an average "payoff" of perhaps 4-6 intelligence score points. In the end, one might interpret Light and Smith's article as presenting evidence that asserted that whatever the difference is in intelligence scores between Caucasians and Negroes, it is surely considerably less than 15 points.

Stinchcombe (1969) made several points about the effect of environment, perhaps the most valuable point being in reference to Head Start. He wrote that insofar as deficits are cumulative, having an enriched experience for only one or two years is not nearly the same as living in an enriched environment for all of a person's first 20 years. The research on the estimates of heritability in the various studies of twins was reviewed by Fehr (1969). Using a different analysis, a lower estimate of the heritability coefficient was found.

Perhaps one of the most cogent discussions of the limitations of Jensen's study was made by Deutsch (1969). As Deutsch and Jensen formerly collaborated in earlier research, Deutsch might be seen as having written a letter to a former colleague asking him to ". . . retract his genetic conclusions in the light of data about and understanding of environmental factors with which he

was apparently not familiar at the time he wrote his article" (p. 552). Deutsch also pointed out that lawyers in some desegregation cases and some legislators have used Jensen's article either to avoid full integration or to underfund public education. In the time since Deutsch made his plea, Jensen has not only *not* retracted his argument but, if anything, is making stronger statements in regard to the importance of heredity.

Several more caustic publications have answered in criticism of Jensen's article. Richardson and Spears (1972) edited a book of essays refuting Jensen, including a particularly scathing essay by Swift, who views Jensen as a heretic from the scientific community. Gartner, Greer, and Riessman (1974) also edited a book of essays, called *The New Assault on Equality*, which a neutral observer might view as leaning more toward a polemic than shedding any new understanding of the controversy.

In 1974, broadening the controversy, Chomsky took Herrnstein to task, stirring up his own hornet's nest. Chomsky's position might be interpreted as requiring the suppression and abandonment of research if it led to findings unpleasant to the egalitarian point of view:

> Turning to the question of race and intelligence, we grant too much to the contemporary investigator of this question when we see him as faced with a conflict of values: scientific curiosity versus social consequences. Given the virtual certainty that even the undertaking of the inquiry will reinforce some of the most despicable features of our society, the intensity of the presumed moral dilemma depends critically on the scientific significance of the issue that he is choosing to investigate. Even if the scientific significance were immense, we should certainly question the seriousness of the dilemma, given the likely social consequences. But if the scientific interest of any possible finding is slight, then the dilemma vanishes.
>
> In fact, it seems that the question of the relation, if any, between race and intelligence has little scientific importance (as it has no social importance, except under the assumptions of a racist society). A possible correlation between mean IQ and skin color is of not greater scientific interest than a correlation between any two other arbitrarily selected traits, say, mean height and color of eyes. (p. 99)

Banesh Hoffman, in his interview by Houts in 1975, had some interesting views regarding the question brought up by Chomsky:

> Houts: Let me ask another question. Do you think that it's morally defensible to inquire into such things as social characteristics?

> Hoffman: Yes, I think so. I think that all sorts of things should be inquired into. I don't think you should set limits on inquiry, although I shouldn't say you should never set limits. . . .But in the matter of social characteristics, I think we have to investigate it very, very carefully. . . .But if your motivation is honestly and seriously scientific, I think that the more we know, on the whole, the better off we are. (p. 37)

I have two arguments with Chomsky's reasoning. First, the denial of access to research, whatever the area, seems defensible to me on only one ground, that direct harm is done to the subjects in the process of carrying out the research. Second, it *is* of social significance to conduct studies regarding individual and group differences. While it may never have been planned for such use, the data available on various groups have been instrumental in pointing out specific instances of social injustice, whether it has involved blacks, Indians, females, males, Chicanos, or any other group.

Finally, Jencks, as I mentioned earlier, has written a provocative book that enters into the discussion. He found, upon reanalysis of Coleman's data, that differences in quality of schooling have little relationship to adult success; he reasoned that, since the quality of schooling may not matter so much, schools should be made a "fun" place to be. He also found that familial incomes (that is, brothers and sisters) varied almost as much as incomes in general. It was also his conclusion that 45 percent (not 80 percent) of intelligence scores were due to heredity. In the end, Jencks seems to have been placed in the Jensen camp by the anti-Jensens, but ignored by the Jensen camp.

What, then, might be made of the heredity-environment issue as it relates to human intelligence? In fairness to Jensen, he has meticulously followed standard practice in both relating his own research and relating the research of others. Surely, much fault can be found not so much with the research but with the dramatic way he attempted to refute compensatory education. More important, it should be pointed out that Jensen never intended to test the heredity-environment question as an impartial judge; by utilizing the genetics approach and attempting to estimate h^2, the heritability index, the whole focus is on heritability. Further, $1-h^2$ is not to be interpreted as environmental variance, since, as Jensen states, "In biometrical genetics, the environmental variance of σ_E^2 is simply the residual non-genetic variance, nothing more." (1975, p. 173) While Jensen never explicitly cautions against the interpretation of his data as a test of the heredity-environment issue, his methodology clearly shows him to be studying heredity, not the latter.

Perhaps psychometric tests have reached their

plateau in regard to the environment-heredity controversy. Clearly, better measurements are needed for the concepts of "heredity" and "environment"; nor is it likely that psychometric devices will provide them. Genetic studies would seem to be more up to that task.

2

A Genetic Approach to Intellectual Inheritance

A second, and potentially more fruitful, approach to hereditary differences in intelligence is through a genetic, as opposed to a totally psychometric, approach to studying individual and group differences. One such study that might be seen as a rudimentary beginning was conducted by Lehrke (1972a).

Lehrke reviewed several studies dealing with mental retardation. One continuing aspect of mental retardation is that, in almost every reported survey, the males predominate over females. Lehrke hypothesized that those genes that are a major determiner of intelligence are located on the X-chromosome. X-linkage of such traits would account for the apparent greater male variability. This greater variability would result in proportionately higher incidence of both mental retardation and of higher intelligence in males. Deleterious alleles of these genes could result in mental retardation that is transmitted as a sex-linked recessive.

Lehrke interpreted the data available as supporting four hypotheses:

> Hypothesis 1. There are major genetic loci relating to human intellectual functioning that are located on the X chromosome.
>
> Hypothesis 2. These genes, if mutated, can lead to subnormal intellectual functioning, including mental retardation, in a X-linked manner.
>
> Hypothesis 3. One or more of these genes relates to verbal functioning.
>
> Hypothesis 4. The deficit relates primarily to the central nervous system. (p. 612)

In support of his hypotheses, he quoted several sources that indicate the greater variability of male intelligence scores. Also, Reed and Reed (1965) are quoted regarding the incidence of known retardates: If the mother but not the father is retarded, the probability of retardation in the children is twice as great as in families where the father is retarded but the mother is not.

Anastasi (1972) pointed out that insufficient evi-

dence is available to talk usefully about a "theory" of X-linkage to major intellectual traits. She indicated that the hypothesis would be usefully tested by investigating several retarded individuals. Nance and Engle (1972) suggested that social reasons may exist for the predominance of males being regarded as retarded. They also stated, in support of Lehrke, that "There is no doubt that there are many sex-linked recessive traits that are associated with mental retardation, and that in the aggregate these syndromes may constitute a greater proportion of the retarded population than most people are aware." (p. 625)

In response to Anastasi and Nance and Engle, Lehrke (1972b) clarified several points that they had raised. While no further rejoinders were included in this particular issue of the *American Journal of Mental Deficiency*, there could be more to Nance and Engle's comment regarding how social mores serve as a basis for decisions about institutionalizing retardates than Lehrke is willing to credit. Unfortunately, documenting evidence of differential bases for admission to institutions for mental retardation is quite difficult, and perhaps impossible in regard to making definitive judgments. In my limited experience in six month's employment as an attendant in a state mental institution, there *seemed* to be a behavior difference between male and female patients. It appeared that females are much more reluctantly consigned to mental hospitals than is true for males. If this tendency holds for other types of institutionalization, including mental retardation, then differences in proportions of males and females in institutions for the mentally retarded may well have a social base.

SEX DIFFERENCES IN MATHEMATICAL TALENT

Bearing on Lehrke's hypotheses as it relates to higher intellectual functioning is the research being conducted at Johns Hopkins University under the direction of Julian Stanley. In a large testing process preliminary to their major study, Keating and Stanley (1972) found an unexpected and disconcerting sex difference showing a preponderance of males with higher mathematical talent. Using the *Sequential Aptitude Test Mathematical* (SAT-M) as a screening device with 396 students participating in a mathematics contest (223 males, 173 females), they found that those scoring 610 or over numbered 43 males but no females. On another test administered at the same time, *Mathematics Achievement Level I* (M-I), 22 males and no females scored 560 or above.

Astin (1974) sought to investigate these sex differences further. Both she and Anastasi (1974) point out that mathematics scores correlate substantially with personality characteristics most readily associated with masculinity (independence, nonconformity, and unconventionality). Within female groups, these correlations

also hold. Astin proffered several social explanations for the results: (1) Girls may become more anxious about competing than boys; (2) role stereotyping favors interests in mathematics and science for boys; and (3) whereas girls appeared to like school, higher scoring males reported a strong dislike for school, showing them to be more nonconforming. While Astin did concede that a biological explanation may account for sex differences in scores, she feels the evidence is more convincing in the direction of cultural reinforcement of differences due to sex-role identification.

In opting for a more parsimonious explanation, Lehrke might argue that gene differences by sex are also related to mathematical talent. Stafford (1961) showed that spatial visualization was sex-linked, with males scoring higher. Hartlage (1970) also investigated this same phenomenon with similar results. Bock and Kolakowski (1973) followed up these two studies with additional data of their own. They found that in all three studies on spatial-visualization the correlations for father-daughter and mother-son were both higher than either father-son or mother-daughter. In all three studies, males scored significantly higher than females. Of additional interest was the finding that spatial-visualization appears to be class-free, in the sense that lower class subjects score as well as middle or upper class subjects. The explanation offered by Bock and Kolakowski is that spatial ability is substantially influenced by a recessive sex-linked gene.

As a final comment, it is fair to say that studies regarding genetic components to intelligence from a biological viewpoint are at a beginning stage, and are by no means definitive in their results. On the other hand, psychometric investigations, wherein all measures used follow a paper-pencil or question-answer format, are less likely, over the long run, to be as convincing as a research viewpoint that includes a biological examination of the heredity component.

3

Looking at Tests and Items on Tests: Hoffman Revisited

The stock in trade of the test sellers is the multiple-choice question. Students who call multiple-choice tests "multiple-guess" tests are closer to the mark than they realize. Several writers have expounded on specific items on the various standardized tests. In the July/August 1975 issue of *Principal*, Zacharias, Schwartz, Cole, and Butler criticized specific items. By far the most interesting exposure of tests and test makers was done by Hoffman (1962).

While the following item does not appear on any published test that I am aware of, it is suggestive of the dilemma that faces test takers:

America was discovered by

(a) Christopher Columbus
(b) Leif Erickson
(c) Welsh sailors
(d) the Chinese
(e) Indians

If the student is told to pick the answer that "is most correct," a knowledgeable reader would indeed be perplexed. There is at least some archeological evidence that would allow answers (a), (b), (c), and (d) to be correct; on the other hand, it is clear that the Indians were already here to greet whomever arrived. Newer archeological evidence may show that even the Indians were preceded by some other group. So the problem here is being clear about what is meant by the word "discovered." Assume for the moment that the keyed answer is (a). Interestingly, high scorers might tend to get this item "right" more often than lower scorers, giving this item a high "discriminant validity," and perhaps permit it to be a survivor of an "item analysis." Higher scorers might use a strategy somewhat like this: the item clearly does not have a right answer, so which answer might appeal to a person who writes tests? While answers (a), (b), (c), and (d) might have sufficient evidence for them to be "discoverers," (b), (c), and (d) did not establish their evidence widely to Caucasian Europeans. Also, the test writer probably uses a Caucasian point of view toward the term "discover," so that answer (e) would be inadvisable; even though the Indians obviously

knew about America, they didn't tell *us*. In other words, high test scorers may have learned how to "get into the heads" of people who write questions.

Sequences of numbers are popular on several aptitude tests. Consider the following item (Otis and Lenon, 1967, p. 2)

What term is missing in this series?

3, 5, 7, ?, 11, 13 (8, 9, 10, 14, 15)

The elicited answer is 9, but mathematically any of the answers can be shown to be correct. The answer 9 could be achieved by the following series: $2n + 1$. The answer 10 could be achieved by using: $2n + 1 + (n-1)(n-2)(n-3)(n-5)(n-6)/12$. To get 8, a similar series can be used: $2n + 1 - (n-1)(n-2)(n-3)(n-5)(n-6)/12$. To get 14, the following series will work: $2n + 1 + 5(n-1)(n-2)(n-3)(n-5)(n-6)/12$. Finally, to get 15, the following may be used: $2n + 1 + (n-1)(n-2)(n-3)(n-4)(n-6)/2$.

A worrisome question may enter the mind of the bright student: while mathematically any of the answers can be correct, does the person who made out the test know this? If the student goes on the assumption that the test is really rather simpleminded (as, perhaps, is the test constructor), then the easiest answer is 9, a piece of reasoning that raises a critical issue almost never addressed either by test constructors or test critics. How can a person who is more intelligent, or more knowledgeable, than the test writer and/or administrator be fairly tested by the test and/or administrator? In the absence of flexibility on the part of the person who scores a test, it may not always be possible. As an interesting aside, I took the WAIS from an examiner in 1964. One of the questions on the WAIS is: "What is the population of the United States?" The answer given, 192 million while closely approximating the actual *counted* population, was technically incorrect going by the test manual, written in 1955. Full credit was given for answers between 140 and 180 million. The test manual was clearly outdated. Further, it is well known that the official population clearly underestimates lower income minority males between the ages of 18-30. The official estimate could be off by at least a few million. Knowledge of this phenomenon may be costly if the examiner is not flexible. (The examiner in question *was* flexible. Later he said, "If you said it's 192 million, then it's probably 192 million. The test manual is outdated anyway.")

What happens when a person taking a test recognizes an error in the instrument? If the experience of the students reported by Hoffman is typical, writing to a test company is of little avail. Apparently they rarely respond to individuals writing to question certain items; at least they rarely admit to error. While item analyses are not always particularly useful, if a "distrac-

tor" or plausible incorrect answer is picked more often than the keyed answer, it is not wholly impossible that a mistake has been made in the key.

ON DISCRIMINANT VALIDITY, ITEM ANALYSIS, AND REFINING PSYCHOMETRIC DEVICES

Standardized tests typically go through a complex item tryout period. One step includes finding the discriminant validity of an item. Typically, the percentage of people whose total score places them in the upper fourth are compared on each item to those whose total score places them in the lower fourth. Suppose 80 percent of those in the upper fourth get an item correct, but only 40 percent of those in the lower fourth get the item correct. The discriminant validity is then .80 - .40 = .40, a fairly respectable discriminant validity. An assumption made is that people who get higher scores know more about the subject than those who get lower scores. What happens if the reverse is true: that is, lower scorers are more likely to get the item correct than those with higher scores? The item is likely to be dropped or rewritten; even if a casual (i.e., non-psychometric expert) reader thinks it seems like a good item, he is still unlikely to convince the psychometric expert of the validity of the item.

Various other technical details go into constructing a psychometric device, including finding measures of reliability and validity, continual item tryout, etc. The non-expert should be advised that the words "reliability" and "validity," in this context, lose their usual meanings, and instead take on technical aspects that, beyond being bewildering, are highly misleading.* Measuring reliability is similar to measuring intelligence. Neither can be measured directly, so that hypothetical constructs must suffice.

After a test has been refined through item analysis, a final form of the test is developed. The refining process is akin to processing wheat: the original product, with minimum processing, might have been more fit for human consumption.

*See, also, an article directly related to the misleading nature of terms in testing by Lazarus, 1975b.

4

Use of Intelligence Testing in the Schools

Intelligence testing has a lengthy history of use in the public schools, but its use has never been more forcefully challenged than it is today. As mentioned earlier, in a recent issue of *Principal*, Zacharias, Morrison, Purvin, Padilla and Garza, and Lazarus have in various ways asked for either a moratorium or the abolition of intelligence testing. In several large cities, including New York, Los Angeles, and Washington, intelligence testing has, in fact, been abandoned on a district-wide scale.

The reasons usually given in support of a moratorium on such tests are either that they reflect an Eastern, Caucasian, middle class bias and discriminate against minority groups, or that they are often used for labeling children, particularly as they are related to placing them in special classes. In addition, teachers and administrators who have access to the scores of these tests are also presumed to be guilty of setting in motion a "self-fulfilling prophecy" about the children. While I don't necessarily disagree with these arguments, I think some distinctions have to be made.

Much of the abuse of intelligence testing is due, I submit, to their uses in a negative sense. If legislation were passed so that intelligence tests could be used only in a positive sense, then perhaps at least some of those who wish either a moratorium or abolition would be willing to concede a real value in intelligence testing. Consider the alternative of abandoning standardized tests in general and intelligence tests specifically. The stage would be set for *status quo* oriented teachers and administrators to more intensively act out their biases and reward conforming students.

An incident that illustrates the point I am trying to make occurred several years ago in a Northern California town. Because a borderline normal female was also quite promiscuous, school leaders tried to get her placed in an institution for the retarded. She was administered a bevy of individualized tests. Had she scored below 70, the school people could have proceeded with the placement. Curiously, her scores on these tests varied between 71-75, never below 70. To me, the incident points out several things. The school officials tried to use the intelligence tests in a negative sense; they were offended by the girl's behavior and

wanted to remove her from the community under the guise that she was retarded. The tests, however, were used in a positive sense; the girl's scores having exceeded the minimum, school officials could not remove her from the community. Had they had a free hand, they would have undoubtedly sent her to an institution for the mentally retarded just to get her out of their community.

Likewise, intelligence tests have been said to discriminate against minorities, a charge that has considerable substantiation. Consider, however, a black student who achieves a score of 140 on the WAIS; that score, in itself, would seem *prima facie* evidence of superior intelligence, regardless of the opinions held of this student by the student's teachers. To discontinue intelligence testing would seem to be a discriminating act against bright minority students. If the rights of a child are to be considered, then the child's right to take a test is at least as great as the child's right not to take a test.

The statement in the previous paragraph is predicated on there being some positive contingencies connected to taking a test. For example, schools might have made special provisions for students who score above a particular score. If, on the other hand, the only contingencies of testing are negative, or if the tests are given purely for administrative reasons, then I would not try to defend the testing process.

The legislation I have in mind might require a school district to forego placing a child in a special education classroom if he achieved above a given score. Nor would the school be allowed to place the child in the special classroom simply because he had failed to achieve the specified score. If the interests of the child were paramount and clearly supported by law, a varied group of professionals--including, perhaps, a physician, a school counselor, a clinical psychologist, the teacher, and a speech therapist--might be called on to assess aspects of the child's growth, in consultation with the parents and the child. This group would then be in a position to make a more judicious choice of helping the child toward maximizing his learning opportunities. If this process seems to the reader to be a lottery where you win sometimes, but never lose, consider that the purpose of public education is the provision of meaningful learning experiences for our nation's youth, *not* the employment of middle class bureaucratic functionaries. Besides, what is wrong with allowing more students the experience of winning?

A point should be made here about the history of intelligence testing. As Kamin (1975) has pointed out, this history includes some unsavory forefathers. While Binet was an important exception, many of the other leaders in the development of intelligence testing (including Terman, Yerkes and Goddard) were proponents of a eugenics view, following the political minds of the day (1912-1925). Concurrent with the construction of early

intelligence tests, Spearman developed the concept of factor analysis, principally as an aid to understanding intelligence. Large-scale intelligence testing was conducted on new recruits during World War I, using the "alpha" test. Intelligence testing was thus given several incentives for advancement despite the bias of several of the early developers. In the process, intelligence testing has received the greatest attention of the various psychometric devices used, and as such, the psychometric qualities of the more widely distributed intelligence tests are perhaps the most advanced examples in the art of psychometric testing. If intelligence tests are of dubious value, then where does this place other psychometric testing, such as achievement testing, personality testing or attitudinal testing? If intelligence tests are to be eliminated, then few tests of any type would seem to be free from a successful challenge from some interest group.

It seems bothersome to many critics of intelligence testing that the concept of intelligence is not "pinned down" to an acceptable concrete entity, but rather is to some degree seen as an abstraction (or hypothetical construct) that is measured indirectly through the various tests, however inadequate. However bothersome the use of an abstraction is to the critics of intelligence testing, it should be pointed out that, with few (if any) exceptions, other personality and/or attitudinal measures are even less well anchored to a concrete reality than is intelligence. Those who would argue against the use of intelligence tests because they don't measure (directly) "intelligence" would seem also to argue against the use of attitudinal tests, such as the measure of authoritarianism devised by Adorno.

Adorno and others (1950) instituted an F scale to test for rigidity of thinking and agreement with an ultra-right (fascistic) viewpoint. One logical criticism that has been raised is that this kind of test is insensitive to discovering members of the ultra-left who are, if anything, even more rigid; Tapp (1975) has termed these people "totalitarian liberals." Rather than discard the F scale because it has been shown to be inadequate in some circumstances, it makes more sense empirically to note its limitations, however numerous, but to allow its continued use for the sake of eventually better understanding the human condition.

Should intelligence testing be abandoned? While any testing can lead to abuse, their positive use seems too important to simply eliminate them. In the long run, society would be the loser.

5

Use of Standardized Achievement Tests in the Schools

Consider for a moment an all-too-frequent type of testing abuse. A school district with perhaps 10,000 students administers to each student each year a complete test battery such as the *Iowa Tests of Basic Skills* or the *California Achievement Tests*. If the district employs 500 teachers, then the 500 teachers spend the equivalent of a full professional day each administering the tests, and the 10,000 students spend the same time taking the tests. Consider the costs. First, the test forms and answer sheets have to be purchased. This outlay could easily run several thousand dollars. The tests have to be scored; if the teachers have to do the scoring, they are likely to spend another 20 to 40 hours in the drudgery of test correcting. If the tests are scored by optical scanner, the cost is likely to run at least a few thousand dollars. The teacher's time, if prorated at $50 a day, comes to $25,000. If the students' time is worth anything (which it very well should be), the students have collectively spent 60,000 to 80,000 hours taking tests. At $2 an hour, this "expense" runs at least $125,000. Even if this "cost" is disregarded, as is generally the case, other student costs should be understood. Many students can see no relevance to spending so much time taking tests. Further, a negative affect toward test-taking is easily envisioned in this massive testing effort. Giving the tests (as they are usually administered) in massive doses, fatigue must surely play a part in students' scores. Considering the teachers again, the boredom that ensues from administering tests hour after hour is taxing; if the teacher is highly motivated to enhance a child's learning experiences, administering standardized tests is extremely frustrating.

What are the "payoffs" from this massive testing effort? In the best possible world, the school district employs 20 to 25 testing-subject matter specialists, who are able to translate test results into remedial and/or accelerated programs for the individual student in continuous consultation with the classroom teacher. These testing-subject matter experts do not now exist in any number sufficient for present testing programs; their non-existence is in some measure due to the school district's unwillingness to expend monies for the necessary number of specialists to implement such a program.

More likely, the school district will use the

scores to arrive at some normative data on comparative achievement levels; if the data is embarrassing, suppression of the information is all too common.

Teachers are not now likely to look closely at achievement test scores. All too often, the test scores sit in files and are only occasionally consulted, if a parent or counselor or other agency insists on the information. Schools rarely embark upon any systematic usage of test information to enhance student learning experiences.

If the costs and probable usage are compared, there is no justification for massive standardized testing. If finding normative data to compare the school district to a national scale is the desire, then it can be achieved without this enormous human cost. If students were randomly chosen, 50 each at grade levels 3, 5, 7, 9, and 11, so that a total of 250 students would be tested, the same usage could ensue and at a greatly reduced cost, both in actual dollars expended and in wasted time. If a moratorium is in order, that moratorium *should* consider mindless massive testing programs that exhibit little payoff for the costs involved.

In fairness to the test constructors (of at least the *Iowa Tests of Basic Skills*), it should be said that the tests themselves are often technically sound; properly used (with a *sample* of the students rather than the whole population), they may yield useful information. Improper use, like overeating, has undesirable side effects.

6

Why Do Schools Administer Standardized Tests?

Why do schools administer standardized tests? If the question is asked of students, a plethora of answers may be given. One answer that probably does *not* occur very often is, "To help us where we have problems." If a teacher is asked this question, a likely response is, "My principal forces all teachers to give them." Central office personnel may either resort to the "district policy" answer, or perhaps indicate that the "public" expects the schools to be accountable. A most *unlikely* answer is, "Because it enhances the learning activities in all classrooms." If Perrone's (1975) experience is typical (and there seems to be no contrary evidence), then the scenario just described may be apt:

> As this eventful week came to a close, I called an assistant principal at a junior high school in yet another school district, opening the conversation with, "How are things going?" His response was not what I had anticipated.
>
> "Terrible!" he exclaimed, "We gave the _____ test on Monday and Tuesday, and the school year hasn't settled down yet. It's like this every school year."
>
> "Why do you give the test?" I asked, but I knew why: it was a systemwide activity. Then I asked what was done with the test results, and the assistant principal said, "We just file them away; no one looks at them." (Perrone, 1975, p. 97)

In other words, for so many, the reason for giving standardized tests this year is because they gave them last year, and for every previous school year in anyone's memory. To discontinue the standardized testing program (if this is seen as a reasoned solution) would mean rocking the boat--and rocking the boat is almost an extinct behavior among school administrators.

But how should the schools replace the standardized tests? First of all, it should free up the time previously used for testing for the teachers to use in more meaningful activities with the students. If the money allocated for test-giving cannot be shifted to other previously neglected areas, then perhaps some tax reductions might occur.

If massive standardized testing is to be abandoned

in a given school district, it should be pointed out that the process of abandonment is perhaps even more important than the outcome. Serious and reasonable discussions at all levels are necessary. Questions have to be raised, such as: "Why do we give (take) these tests?"; "How might the test scores be meaningfully employed by the student, the teacher, the parent, the administrator, or the researcher?"' "*Will* we meaningfully use the test?"; "Is the test worth what it costs (in dollars and in time)?"; "If we have a standardized testing program, is it mandatory, or could classes (students) elect to participate or not to participate?"; "Let's look at the tests; do they test relevant areas?"; "Are the tests the most economical way to sample necessary instructional areas, or are other testing forms (such as problem solving) more relevant?"

If all who are involved with the testing area *reason* together, then whatever their reasoned decision, that they have wrestled with the necessary questions and arrived at a decision is sufficient. Even if they decide to continue the standardized testing program in some form, they are more likely to understand the rationale for testing, and are more likely to make use of the results.

7

Do We Need This Big an Educational Testing Industry?

The dimensions of the educational testing industry are immense; Kohn (1975) has estimated its yearly volume at $150 million. Corporations such as the Education Testing Service (ETS), Psychological Corporation, American College Testing Service, and the California Test Bureau of McGraw-Hill are fairly widely known. Textbook publishers are also heavily invested in the testing industry; Houghton-Mifflin, for one, publishes the *Iowa Tests of Basic Skills*, the *Lorge-Thorndike Intelligence Tests*, and the *Stanford-Binet Intelligence Scale*. These corporations are but the tip of the iceberg. There is a startling array of published tests available to test every conceivable aspect of human endeavor. Buros (1974) lists more than 100 published group intelligence tests; more than 20 individual tests are available. The total number of published tests available easily numbers in the thousands (Buros, 1972, lists 2,585 tests).

More important than the number of tests is the effect testing has on human lives. Students typically have to take (and pay for) a college admissions test even while they are in high school. Companies such as ETS have developed a whole battery of achievement tests that are now accepted for college credit if the student scores high enough. It is now possible for a student to achieve one year's credit at a designated college or university in one day's effort taking ETS's advanced placement tests before the student has even set foot on the campus.

If the student wants to go to graduate school, then ETS's Graduate Record Examinations (GRE) are usually required. And pity the student in a competitive academic area who may have excellent undergraduate grades and excellent recommendations but only mediocre GRE's. Similarly, there are tests for almost every professional school, but surely including law and medicine.

If the student needs a loan, again ETS comes to the rescue (at $3.75, thank you). The student fills out an optical scanner sheet, and somehow this is better than showing need to local officials.

Suppose the student has to show a reading knowledge of a foreign language for the Ph.D. Do the professors in the appropriate foreign language test the student? Not on your life. ETS comes to the rescue again.

One of the funnier incidents in my professional experience is due to the use of the ETS French Examina-

tion. The graduate school where I taught decreed that a student should score at the 33rd percentile (or higher) to show a reading knowledge of the language. There were three forms of the test, one for students in the sciences, one for students in the humanities, and one for students in the social sciences. However, the first part of all three tests was identical. Because students in the humanities and social sciences naturally have a greater language background, their scores are usually higher. Thus getting a score of the 33rd percentile is much harder for students in these areas than it is for those in the sciences. As it happened, one could pass the science test by doing only the grammar section and omitting the rest of the test; such was not possible for the other two tests. In fact, several students in the humanities and social sciences section "failed" the language examination even though they had obtained more than a hundred points in excess of the score required to pass the science examination. As an alternative strategy, I suggested to several affected students that they take the science examination and do the best they could on the grammar section. "If you've done well enough," I said, "you don't even have to look at the science reading section of the test." By hook or crook, students started passing the language examination significantly more often.

One additional aspect of the testing industry that should be considered is its secrecy. Kohn (1975), in an attempt to gather information about the testing industry, talked to representatives from the major testing firms. None of the people he talked to was willing or able to estimate the number of students tested each year in the United States. They also would not reveal how many tests their individual companies had sold or scored. The only comment made in regard to overall size of the testing industry intimated that it was too *small*, and should necessarily become *larger*.

A short reflection on the amount of autonomy given to the industry by the colleges and universities is nothing short of astonishing. One wonders what will be the next coup that the testing industry will make. Are they really very far from offering degrees? It has been said, in fact, and only half in jest, that the largest university in the world is none other than ETS. Surely, more college credits are achieved by taking ETS tests than any one university offers.

One of the more eloquent arguments against the testing industry was made by Karier (1972). In it, he indicated the testing industry as being the servant of power, privilege, and status. A parallel can be drawn with the large philanthropic foundations. The role of the foundations (Carnegie, Ford, and Rockefeller), as a "fourth" branch of government, is said to be that of a manipulator of educational thinking for the corporate liberal state. The projects of the foundations have been a pressure valve for society; studies ultimately "prove" the system is working, but needs some adjustment. As an

example of how the corporate state has attempted to influence the direction of public education through massive testing, Karier cites evidence regarding the testing done with the national assessment program. The establishment of the national assessment is seen as a very definite step toward a national curriculum.

A similar view is given by Kamin (1974, 1975). Kamin reviews the history of the development of intelligence testing and notes that, with the exception of Binet, many of the earlier developers of intelligence testing had a decidedly hereditarian view of intelligence. This hereditarian view was, in turn, made to serve the politics of those who wanted to impose differential immigration quotas on the so-called "genetically inferior" peoples of the South Mediterranean and Eastern European areas.

Should so large and intrusive a testing industry be encouraged to continue? Surely the testing bureaus would answer not only in the affirmative, but with reasons for why the testing industry should and will get larger. But another ascendant point of view, a point of view of students and faculty, is that we don't need one-tenth of the "services" testing firms furnish. Many of the services do no more than help enforce the *status quo*. What possibly could be more indicative of enforced submissiveness than for a third grade child to sit for six hours taking tests whose purpose no one seems able to understand? The teacher doesn't have the authority to excuse the child; the principal may also feel the lack of authority. Higher up the chain, they are so far removed that the anguish of the individual child is quietly laughed off. What function do such experiences have in a democratic society?

If the testing industry grew "just like Topsy," then is it not time to greatly reduce the size of the weed?

8

An Interim Solution: Teaching Students to Take Tests

Even if those in the testing industry, on their own, were suddenly to reverse their growth and retrench to a former shape, it is likely that their tests would be with us for a long time to come. In any case, what's more probable is that the immediate future will bring continuing growth. Thus, students should be prepared for the testing onslaught.

One example of preparing students for tests was given in the previous chapter. Perhaps where such coaching is most necessary is with younger students who have to endure the *Iowa Tests of Basic Skills* or the *California Achievement Tests*. I am not implying coaching in the illegal sense, as in the Turnkey project in Texarkana in the early 1970s, where a large-scale performance-contracting experiment was accused of using actual test items in their instruction prior to the scheduled testing. Rather, preparation, much as a baseball coach gets a team ready for a game, is helpful. Students should not have to encounter a new testing form when the tests are taken. That is somewhat like a Little League player going to bat, striking out without taking the bat off his shoulders, and then remarking, "Gee, I never saw a left-handed pitcher before!" Students should be allowed some experience with the multiple-choice format and its ludicrous idiosyncracies. They should know about time limits and whether or not there is a penalty for guessing. If there is no penalty for guessing, then an answer should be given for every item, even if the question has not been read. And even if there is a penalty, the child should know that he is better off (in terms of the score) if a guess is made, if anything at all is known about the area being tested. As an example, suppose a question has four answers and the child knows that one of the answers is wrong, but the other three are possibly correct. Because the likely penalty for guessing is C - 1/4I (Correct - 1/4 X Incorrect), knowing that one answer is incorrect moves the child's probability of getting a positive score on this item above zero: probability = 1/3 - 1/4 = 1/12. While this value is only slightly higher than zero, because it is above zero, the score is maximized if a guess is made.

Some of the coaching I have in mind even makes me cringe; however, the point is this: if students are to maximize their scores, they do need some battlefield ex-

perience. Thus, the teacher is encouraged to write some ambiguous items similar to those found on tests and let the students cut their teeth on a miniature version of "multiple guess"; this is somewhat akin to telling the players to get ready for some bad umpiring.

If these suggestions seem unrelated to a child's educational experience, consider that the same could be said for the testing. The suggestions clearly have nothing to do with actual learning; their value is survival in the testing jungle.

But is is important, too, to go beyond the negative aspects of survivorship to where positive recommendations can be considered for creating a clearing in that jungle. School districts, for one thing, should be held accountable for their testing programs. Perhaps some of the tests can pass the acid test; many would not. Surely, district-wide massive testing would serve as a useful target for accountability and possible litigation. A school district should have to show a ledger giving the costs and benefits for the various tests used. Costs must surely include teacher and student time. For those tests that do not have a favorable payoff relative to the cost, changes should be mandated.

Second, it is clear that district-wide use of individualized intelligence tests is too expensive in time and money; giving the tests on a selected basis may be useful, however, Suppose a black male child is being considered for special education on the basis of teacher recommendations, test scores, and whatever else goes into the decision. The child should have the right to take an intelligence test under circumstances that maximize his score. For example, a black, male examiner who is able to elicit the child's best effort would be a minimum. The test (or tests) chosen should as nearly as possible reflect experiences that would normally be available to him. If under these maximized conditions the child does in fact "pass" the test, the learning experiences should be structured so that the child might be integrated as quickly as possible into a "regular" classroom setting.

One other word of advice should be followed by anyone who has any contact with any kind of test score. Before interpreting any score, the person should first take the test under the same conditions as do students. If little value is seen in the test, then the same value can be attributed to a test score.

9

A More Lasting Solution: Constructing Locally-Useful Criterion-Referenced Tests

For most learning activities, it would seem that the most logical people to write the tests would be those who are involved in them. Thus teachers, locally-employed test specialists (if any), and students might all have a hand in test construction.

Implicit in the process is a placement of stronger responsibility upon the teachers and students than typically has been given to them. A first step is the decision as to the material to be learned. How is the material important? What are the specific goals that need to be met? After having completed a particular course or unit, what "survival value" has the learning experience? That is, what is it, either in the subject matter or in the method of learning, that will be useful now or later on? How and where will it be useful?

The kinds of questions implied above are somewhat similar to the concerns of those who have emphasized behavioral objectives (such as Mager, 1962) or criterion-referenced tests (such as Popham, 1975); more simply, they are questions that a thinking teacher is likely to ask in regard to a learning experience.

Also, those involved with test construction would be advised to become somewhat more knowledgeable in test construction. While attendance in university courses would sometimes be helpful, test construction courses tend to emphasize multiple-choice items to the neglect of other areas. Multiple-choice items have several drawbacks. First, from the viewpoint of the item writer, it is extremely time-consuming to write multiple-choice tests. Second, and more important, the multiple-choice format is not always a useful manner to test a person's understanding of an area. If, for example, the intent of a course is to teach the students to write short essays, it would appear that the testing situation would have the student write short essays.

While allowing students to "test out" of a course for credit seems justified, it is ironic that universities so readily accept "credits" from ETS and yet often deny students the opportunity to "test out" of a course by taking an examination written by local people. Perhaps universities have not been responsive enough to the needs of students in regard to credit by examination. Given the prod by ETS, those institutions that fail to respond to student need by developing tests locally are

accepting an even bigger intrusion by ETS in their future.

What is the place of standardized testing in such a scheme? If the test exactly fits the goals of the learning experience, then of course the standardized test is appropriate. This might be particularly true in skill courses such as typewriting or in courses based on a national science curriculum.

Perhaps the most notable changes implied are that fewer tests would be given, and the tests that are given can be directly related to actual learning goals. Teachers have been making out tests for almost as long as there have been teachers. If teachers are given appropriate instruction in evaluating the learning activities of students, it would seem that their teaching activities could be enhanced. For that matter, students should be instructed by teachers in the art of making tests. Among the outcomes of attempting to compose such questions, students should be able to better understand their learning.

10

An Alternative Solution: Constructing Criterion-Referenced Work Samples

Perhaps a most useful solution to the testing problem is the construction of criterion-referenced work samples. What, exactly, should students be able to do when they complete a unit or a course? The testing should elicit the same behavior as the stated goal. If the goal is to have the individual student master the use of laboratory equipment, then a multiple-choice test will rarely yield a satisfactory measure of that criterion skill.

One perlexing quality of some standardized tests is that students who have not taken the course supposedly measured by the test do as well or better on the examination as do students who have taken the course. This happens particularly in housekeeping-related courses, such as consumer economics, offered in many high school curricula. Several explanations could be offered, but a rather straightforward one is that the material being tested is learned in many sources besides classrooms. Watching television, reading newspaper columns such as Sylvia Porter's, visiting the local supermarket, reading advertisements, and related activities that are often available to many people beyond those in a course in consumer economics may well be sufficient to demonstrate competency. Rather than trying to devise tests that allow those who take the course to demonstrate higher proficiency, it seems useful to point out that people who do the activities necessary to score high on a test are performing exactly as students might who took a course in consumer economics. The more advisable solution is to accept the time spent watching relevant television programs, reading economics columns, and going to the supermarket (which in educationese is called a field trip) as an alternative learning experience.

AN EXAMPLE OF USING WORK SAMPLES

The supposed reasons for requiring graduate students to take courses in applied statistics is so that the student later will be able to apply what has been learned in their own research. It would seem reasonable then to construct the course in such a way that the student might, as an end product, show competence in executing a research project, even though the dimensions of that project may be necessarily limited due to the con-

straints of time, experience, and the availability of research subjects. In my experience, such constraints have not been nearly as important as I would have expected; students often have better access to researchable data than many of us would guess.

To take the goal one step further, one of the stated reasons for conducting graduate research is the publication (at least in part) of the research conducted. While, on the one hand, this clearly would be an indefensible goal to *require* of all students in a statistics course, on the other hand, using a research journal style has several payoffs in terms of the expectancies of graduate school. To master a research journal style, it is useful first to have read several other articles in one's own field. Because the students are becoming familiar with the state of research in their area of interest, they can personally evaluate the usefulness of their own effort. Also, some students do, in fact, publish their papers eventually.

Note that nothing has been said about typical classroom tests. Indeed, it may be that tests may interfere with the students' and instructor's goals, rather than enhance them. It is probably true that students using the approach just described learn in great detail only those statistical techniques appropriate to their specific project. In many cases, they will be learning statistical techniques not even covered in the main body of the course. But then, isn't that what learning research techniques is all about? Does it not make sense to learn about a technique, in detail, as an application presents itself, whether or not it is required in a formal research course? Such a teaching method is neither particularly new (a previous article appeared in Williams, 1970) nor unique; Novick and Jackson (1974) have also been using a similar technique in teaching Bayesian statistical research methods.

CHOOSING GOALS

Clearly knowing one's goals in teaching or directing any learning experience is a first step toward constructing criterion-referenced work samples. Perhaps helpful in delineating those goals is Mager's (1973) book. Having chosen the goals, the use of a circuitous route through standardized testing would seldom make sense. Surely, tests might be used in a summative sense to ensure that skills have been acquired. But acquiring skills is rarely a sufficient goal in cognitive-oriented learning experiences; rather the utilization of those acquired skills--mastery--seems to make more sense.

A note of caution is in order. Educators should be cognizant of different learning goals even within an extremely homogeneous classroom. Not all students are prepared for a work sample approach to evaluation. In that matching student learning styles with their actual

learning experiences is more important than the utilization of any single teaching strategy, it is well to remember that some students are so oriented to the traditional classroom procedure that, at least for such students, learning experiences in a traditional format (but gradually moving toward a more open format) may maximize their present learning experiences.

11

A Final Statement

What, then, is this third view of educational testing? If the first two views are seen as being nearly diametrically opposed to one another, one advocating a continuation (and perhaps an increase) in the present reliance on standardized testing (and perhaps testing in general), and the second view favoring the discontinuation of standardized and intelligence testing, then the third view advocated in this paper might be seen as being somewhere between those two extremes

The use of standardized tests has tended to make teachers functionaries in the decision-making process regarding the outcomes of the educative process; as the standardized test is seen as "the criterion," innovative educational practices are discouraged unless they can show some competitive edge on "the criterion." This tends to block the development of alternative goals. Rather than help the teacher become a fully functioning professional, and more proficient in measuring student progress, the job of testing is given over to the "experts." That is not to deny that educational testing expertise is a valuable commodity, but it should be more widely diffused among those actually involved in the teaching-learning process. Inherent in this criticism is a plea for alternative ways of measuring goal attainment. However useful the multiple-choice format is, other measures of expressing goal attainment have been neglected to our detriment.

Perhaps the most damning criticism of those in the testing movement is not that they don't have a useful product, but that their product's practical use becomes an abuse. To require every student in a school setting to spend six or eight hours on an examination seems indefensible. The incorrect interpretation of test results, particularly as it applies to intelligence tests, is also indefensible. On the other hand, the abolishment of the testing movement, as is sometimes suggested by its detractors, seems ill-advised. Tests *do* provide us with another perspective. It is not particularly bothersome to hear someone remark about an intelligence test, "In that this student is a bilingual Chicano, I don't think we can say a score of 80 does the student justice." At the same time, it would be an extreme injustice to a student if he scored 130, but had the test disregarded on the basis of its being biased.

If we accept that tests can be biased, and if we accept that people can be biased, then another approach is to understand the nature of the bias. But, in fairness to the student, the information available should be used in such a way that the student's learning potential can be maximized. The construction of newer tests, which either remove the bias or can be used to enhance a student's learning, is helpful. Just as some students don't perform well on some tests, other students don't interact well with some teachers. To label either group on the basis of a limited bit of information is perhaps the worst bias of all.

What should be the fate of the standardized testing movement? The "fate" suggested here is that those who are most involved in the testing process should, in some meaningful way, begin reasoning together on the uses and/or abuses of testing (or for that matter, the uses and abuses of *not* testing). Clearly, no national mandate, whether it favors some type of moratorium or an extension of the testing movement, would allow sufficient individual participation in the reasoning process.

When students ask, "Why do we have to take this test?", it is hoped that the teacher and/or principal can offer sound educational reasons as it relates to the individual's learning experience. If those reasons are not satisfactory and the student is well apprised of the contingencies of not taking the test, then the student could be excused from the testing process. Perhaps as the use of tests becomes an outgrowth of a participatory democracy, wherein the various persons involved reason together, the promise of the testing industry to help enhance students' learning would start to be fulfilled. Under such a system, wherein each person understands and willingly accepts the rationale for the test or tests utilized, testing would be seen (and function) as an integral part of the learning process, rather than as a bulwark to a burgeoning bureaucracy.

References

Adorno, T.W., Frenkel-Brunswick, E.F., Levinson, D.J. and Sanford, R.N. *The authoritarian personality*. New York: Harper & Row, 1950.

Anastasi, A. Commentary on the precocity project. In Stanley, J.C., Keating, D.P. and Fox, L.H. *Mathematical talent*. Baltimore: The Johns Hopkins University Press, 1974, 87-100.

_____. Four hypotheses with a dearth of data: Response to Lehrke's "A theory of X-linkage of major intellectual traits." *American Journal of Mental Deficiency*, 1972, 76, 620-622.

Astin, H. Sex differences in mathematical and scientific precocity. In Stanley, J.C., Keating, D.P. and Fox, L.H. *Mathematical talent*. Baltimore: The Johns Hopkins University Press, 1974, 70-86.

Bane, M.J. *Tests and testing*. McLean, Va.: National Council for Advancement of Educational Writing, 1974.

Bereiter, C. The future of individual differences. *Harvard Educational Review*, 1969, 39, No. 2, 310-318.

Bock, R.D. and Kolakowski, D. Further evidence of sex-linked major-gene influence on human visualizing ability. *American Journal of Human Genetics*, 1973, 25, 1-14.

Brazziel, W.F. A letter from the South. *Harvard Educational Review*, 1969, 39, No. 2, 348-356.

Buros, O.K. (Ed.) *Tests in print II*. Highland Park, N. N.Y.: Gryphon Press, 1972.

_____. *Seventh mental measurements yearbook*. Highland Park, N.Y.: Gryphon Press, 1974.

Burt, C. The inheritance of mental ability. *American Psychologist*, 1958, 13, 1-15.

_____. The genetic determination of differences

in intelligence: A study of monozygotic twins reared together and apart. *British Journal of Psychology*, 1966, 57, 137-153.

Butler, J. Looking backward: Intelligence and testing in the year 2000. *Principal*, 1975, 54, No. 4, 67-75.

Chomsky, N. The fall of Richard Herrnstein's IQ. In Gartner, A., Greer, C. and Riessman, F. *The new assault on inequality*. New York: Perennial Library, 1974

Cohen, J. Multiple regression as a general data-analysis system. *Psychological Bulletin*, 1968, 70, 426-443.

Cole, M. Culture, cognition, and IQ testing. *Principal*, 1975, 54, No. 4, 49-52.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, Modd, A.M. Weinfield, E.D., and York, R.L. *Equality of educational opportunity*. Washington: U.S. Department of Health, Education and Welfare, U.S. Government Printing Office, 1966.

Cronbach, L.J. Heredity, environment, and educational policy. *Harvard Educational Review*, 1969, 39, No. 2, 338-347.

Crow, J.F. Genetic theories and influences: Comments on the value of diversity. *Harvard Educational Review*, 1969, 39, No. 2, 301-309.

Dennis, W. The performance of Hopi Indian children on the Goodenough Draw-a-Man Test. *Journal of Comparative Psychology*, 1942, 34, 341-348.

Deutsch, M. Happenings on the way back to the forum. *Harvard Educational Review*, 1969, No. 3, 523-557.

Ebel, R.L. Educational tests: Valid? Biased? Useful? *Phi Delta Kappan*, 1975, 757, 83-88.

Elkind, D. Piagetian and psychometric interpretations of intelligence. *Harvard Educational Review*, 1969, No. 2, 319-337.

Erlenmeyer-Kimling, L. and Jarvik, L.F. Genetics and intelligence: a review. *Science*, 1963, 142, 1477-1479.

Fehr, F.S. Critique of hereditarian accounts of intelligence and contrary findings: a reply to Jensen. *Harvard Educational Review*, 1969, 39, No. 3, 571-580.

Gartner, A., Greer, C. and Riessman, F. (Eds.) *The new*

assault on inequality. New York: Perennial Library, 1974.

Green, R.L. Comments on Ebel's defense of tests and testing. *Phi Delta Kappan*, 1975, 57, 88-89.

Hartlage, L.C. Sex-linked inheritance of spatial ability. *Perceptual and Motor Skills*, 1970, 31, 610.

Herrnstein, R.J. I.Q. *Atlantic Monthly*, 1971, September, 43-64.

Hoffman, B. *The tyranny of testing*. New York: Crowell-Collier Press, 1962.

_____. Interviewed by Houts, P.L. A conversation with Banesh Hoffman. *Principal*, 1975, 54, No. 6, 30-39.

Hunt, D.E. *Matching models in education*. Ontario, Canada: The Ontario Institute for Studies in Education, 1971.

Hunt, J. McV. Has compensatory education failed? Has it been attempted? *Harvard Educational Review*, 1969, 39, No. 2, 278-300.

Jensen, A.R. How much can we boost IQ and scholastic achievement? *Harvard Educational Review*, 1969, 39, No. 1, 1-123.

_____. The meaning of heritability in the behavioral sciences. *Educational Psychologist*, 1975, II, 171-183.

Jencks, C. *Inequality: a reassessment of family and schooling in America*. New York: Basic Books, 1972.

Kagan, J.S. Inadequate evidence and illogical conclusions. *Harvard Educational Review*, 1969, 39, No. 2, 274-277.

Kamin, L.J. *The science and politics of IQ*. Potomac, Md.: Lawrence Erlbaum Associates, 1974.

_____. The politics of IQ. *Principal*, 1975, 54, No. 4, 15-22.

Karier, C.J. Testing for order and control in the corporate liberal state. *Educational Theory*, 1972, 22, 159-180.

Keating, D.P. and Stanley, J.C. Extreme measures for the exceptionally gifted in mathematics and science. *Educational Researcher*, 1972, 1, No. 9, 3-7.

Kohn, S.D. The numbers game: How the industry operates. *Principal*, 1975, 54, No. 6, 11-23.

Lazarus, M. On the misuse of test data: A second look at Jencks's "Inequality." *Principal*, 1975a, 54, No. 4, 76-78.

_____. Coming to terms with testing. *Principal*, 1975b, 54, No. 6, 24-29.

Lehrke, R. A theory of X-linkage of major intellectual traits. *American Journal of Mental Deficiency*, 1972b, 76, 626-631.

Lewontin, R.C. Race and intelligence. *Bulletin of the Atomic Scientists and Public Affairs*, 1970, 26, No. 3, 2-8.

Light, R.J. and Smith, P.J. Social allocation models of intelligence. *Harvard Educational Review*, 1969, 39, No. 3, 484-510.

Mager, R.F. *Preparing instructional objectives*. Belmont, California: Fearon Publishers, 1962.

_____. *Measuring instructional intent*. Belmont, California: Fearon Publishers, 1973.

McCall, R.B. *Intelligence and heredity*. Homewood, Illinois: Learning Systems Co., 1975.

Morris, F.L. The Jensen hypothesis: Was it the white perspective or white racism? *Journal of Black Studies*, 1972, 2, 371-386.

Morrison, P. The bell shaped pitfall. *Principal*, 1975, 54, No. 4, 34-37.

Nance, W.E. and Engle, E. One X and four hypotheses: Response to Lehrke's "A theory of X-linkage of major intellectual traits." *American Journal of Mental Deficiency*, 1972, 76, 623-625.

Novick, M.R. and Jackson, P.H. *Statistical methods for educational and psychological research*, New York: McGraw-Hill, 1974.

Otis, A.S. and Lenon, R.T. *Otis-Lenon mental ability test, elementary level, Form J*. New York: Harcourt Brace and World, 1967.

Padilla, A.M. and Garza, B.M. A case of cultural myopia. *Principal*, 1975, 54, No. 4, 53-58.

Perrone, V. Alternatives to standardized testing. *Principal*, 1975, 54, No. 6, 96-101.

Popham, W.J. *Educational evaluation.* Englewood Cliffs, N.J. Prentice-Hall, 1975.

Purvin, G. The hidden agendas of IQ. *Principal,* 1975, 54, No. 4, 44-48.

Reed, E.W. and Reed, S.C. *Mental retardation: A family study.* Philadelphia: W.B. Saunders, 1965.

Richardson, K. and Spears, D. (Eds.) *Race and intelligence.* Baltimore: Penguin Books, 1972.

Rosenthal, R. and Jacobson, L. *Pygmalion in the classroom.* New York: Holt, Rinehart and Winston, 1968.

Samuda, R.J. *Psychological testing of American minorities.* New York: Dodd and Mead, 1975.

Schwartz, J.L. The illogic of IQ tests. *Principal,* 1975, 54, No. 4, 38-41.

Shockley, W. Negro IQ deficit: Failure of a malicious coincidence model warrants new research proposals. *Review of Educational Research,* 1971, 41, 227-248.

Sitgreaves, R. Comments on the Jensen report. Paper presented at the meeting of the National Academy of Education, UCLA, October 11, 1969.

Stafford, R.E. Sex differences in spatial visualization as evidence of sex-linked inheritance. *Perceptual Motor Skills,* 1961, 13, 428.

Stinchcombe, A.L. Environment: The cumulation of events. *Harvard Educ nal Review,* 1969, No. 3, 511-522.

Swift, D. What is the environment? in Richardson, K. and Spears, D. (Eds.) *Race and intelligence.* Baltimore: Penguin Books, 1972.

Tapp, J. interviewed by Berman, G. The notion of conspiracy is not tasty to Americans. *Psychology Today,* 1975, 8, No. 12, 60-67.

Wechsler, D. *The measurement and appraisal of adult intelligence.* 4th ed. Baltimore: Williams and Wilkins, 1958.

Williams, J.D. Creativity and the initial statistics course: An exercise in goal oriented behavior. *Journal of Business Education,* 1970, 65, 155-156.

Zacharias, J.R. The trouble with IQ tests. *Principal,* 1975, 54, No. 4, 23-29.

Also available as part of the North Dakota Study Group on Evaluation series:

Observation and Description: An Alternative Methodology for the Investigation of Human Phenomena
Patricia F. Carini

Alternative Evaluation Research Paradigm
Michael Quinn Patton

An Open Education Perspective on Evaluation
George E. Hein

A Handbook on Documentation
Brenda Engel

The Teacher Curriculum Work Center: A Descriptive Study
Sharon Feiman

Deepening the Questions About Change: Developing the Open Corridor Advisory
Lillian Weber (in preparation)

The Word and the Thing: Ways of Seeing the Teacher
Ann Cook and Herb Mack

Special Education: The Meeting of Differences
Steven D. Harlow

Single copies $2, from Vito Perrone, CTL
U. of North Dakota, Grand Forks, N.D. 58202